CNARA: reliability assessment for genomic copy number profiles

Ai, Ni ; Cai, Haoyang ; Solovan, Caius ; Baudis, Michael

Abstract: BACKGROUND DNA copy number profiles from microarray and sequencing experiments sometimes contain wave artefacts which may be introduced during sample preparation and cannot be removed completely by existing preprocessing methods. Besides, large derivative log ratio spread (DLRS) of the probes correlating with poor DNA quality is sometimes observed in genome screening experiments and may lead to unreliable copy number profiles. Depending on the extent of these artefacts and the resulting misidentification of copy number alterations/variations (CNA/CNV), it may be desirable to exclude such samples from analyses or to adapt the downstream data analysis strategy accordingly. RESULTS Here, we propose a method to distinguish reliable genomic copy number profiles from those containing heavy wave artefacts and/or large DLRS. We define four features that adequately summarize the copy number profiles for reliability assessment, and train a classifier on a dataset of 1522 copy number profiles from various microarray platforms. The method can be applied to predict the reliability of copy number profiles irrespective of the underlying microarray platform and may be adapted for those sequencing platforms from which copy number estimates could be computed as a piecewise constant signal. Further details can be found at https://github.com/baudisgroup/CNARA . CONCLUSIONS We have developed a method for the assessment of genomic copy number profiling data, and suggest to apply the method in addition to and after other state-of-the-art noise correction and quality control procedures. CNARA could be instrumental in improving the assessment of data used for genomic data mining experiments and support the reliable functional attribution of copy number aberrations especially in cancer research.

DOI: https://doi.org/10.1186/s12864-016-3074-7

METHOD

CNARA: reliability assessment for genomic copy number profiles

Ni Ai[1*], Haoyang Cai[2], Caius Solovan[3] and Michael Baudis[1*]

[*]Correspondence: ni.ai@uzh.ch; michael.baudis@imls.uzh.ch
[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Full list of author information is available at the end of the article

Abstract

Background
DNA copy number profiles from microarray and sequencing experiments sometimes contain wave artefacts which may be introduced during sample preparation and cannot be removed completely by existing preprocessing methods. Besides, large derivative log ratio spread (DLRS) of the probes correlating with poor DNA quality is sometimes observed in genome screening experiments and may lead to unreliable copy number profiles. Depending on the extent of these artefacts and the resulting misidentification of copy number alterations/variations (CNA/CNV), it may be desirable to exclude such samples from analyses or to adapt the downstream data analysis strategy accordingly.

Results
Here, we propose a method to distinguish reliable genomic copy number profiles from those containing heavy wave artefacts and/or large DLRS. We define four features that adequately summarize the copy number profiles for reliability assessment, and train a classifier on a dataset of 1522 copy number profiles from various microarray platforms. The method can be applied to predict the reliability of copy number profiles irrespective of the underlying microarray platform and may be adapted for those sequencing platforms from which copy number estimates could be computed as a piecewise constant signal. Further details can be found at https://github.com/baudisgroup/CNARA.

Conclusions
We have developed a method for the assessment of genomic copy number profiling data, and suggest to apply the method in addition to and after other state-of-the-art noise correction and quality control procedures. CNARA could be instrumental in improving the assessment of data used for genomic data mining experiments and support the reliable functional attribution of copy number aberrations especially in cancer research.

Keywords: copy number profile; CNA; reliability assessment

Background

Since the introduction of molecular-cytogenetic technologies for whole genome copy number aberration screening [1, 2], considerable advances have been made to work with a variety of sub-optimal material (e.g. micro dissected samples, aspiration biopsies, paraffin embedded tissue), both in the areas of DNA preparation, labeling and platform technologies as well as in bioinformatic processing of the experimental read-out. However, DNA copy number profiles from current microarray and

sequencing experiments sometimes suffer from the presence of systematical "wave patterns" [3] throughout the whole genome, where within each genomic segment the estimated copy number deviates from the true value which is supposed to be a constant. These wave artefacts disrupt the piecewise constant signal of the copy number data and may lead to false positives or negatives in identifying CNAs.

One of the known causes to the wave artefacts is differential DNA retrieval across chromosomal regions, which may be due to GC-content bias [4], DNA replication timing [5], differences in chromatin organization during DNA isolation [6] and damages to the DNA by fixation procedures [7]. Copy number profiles with heavy wave artefacts sometimes can be corrected if certain requirements are met. Marioni et al. developed a method to remove wave artefacts in copy number profiles for normal samples without obvious CNAs [8]. Wiel et al. suggested to eliminate waves in tumor profiles with many CNAs using calibration profiles [3]. Some methods correcting GC-content bias have also been implemented for microarray [4] and sequencing experiments [9]. However, the methods proposed in these studies have a limited ability to remove wave artefacts or put many restrictions on the type and variability of the input data itself. In addition to wave artefacts, large derivative log ratio spread (DLRS) [10] correlating with poor DNA quality also leads to unreliable copy number profiles.

The limited ability of existing experimental and bioinformatic methods to remove wave artefacts or to correct for source dependent DNA quality variations motivated us to devise a method for assessing if genomic copy number can be reliably estimated from a pre-processed, technology agnostic copy number profiling dataset. Rather than developing a method for improving the sample derived copy number profiles themselves, our primary intention here is to provide measures for the contamination of copy number profiles through artefacts and thereby to support decisions regarding the suitability of these copy number profiles for downstream data analysis and interpretation.

Zhang and Zhang previously designed such a measure in an explorative study [7] where they proposed using autocorrelation scanning profile (ASP) to evaluate data quality, and demonstrated on simulated data that the median of ASP (medASP) can be used as a discriminative metric. However, it will be shown in the Discussion section that medASP is not an adequate measure for real-world data of different scenarios.

In this paper, we assess the reliability of copy number profiles using a machine learning approach. From our experience and by experiment, we selected four features which are able to adequately represent the copy number profiles for reliability assessment, i.e. the number of steps that can be detected by the step-fitting algorithm in the copy number profile, a quantitative value indicating how much the copy number data is step-like, the number of segments induced by both CNAs and wave artefacts detected by circular binary segmentation [11, 12], and the mean of DLRS within segments. We will explain in detail about these features in the Results and discussion section and discribe the step-fitting algorithm by which the first two features are generated in the Materials and methods section. Based on these features, we trained a classifier to predict the reliability of the copy number profile. We will describe the way of classifying reliable and unreliable profiles, and subsequently

assigning them into one of the five subcategories each having a biological or experimental correspondence. Our reliability assessment method can also be adapted to assess copy number profiles generated by those sequencing platforms from which copy number estimates could be computed as a piecewise constant signal.

Results

Piecewise constant model and segmentation

Tumor samples often contain CNAs, in which chromosomal segments are found gained or lost in copy number, deviating from the normal diploid status. Genome-wide copy number can be depicted as a piecewise constant signal, where the change-points are the boundaries of the chromosomal segments that differ in copy number, and the constant value between a pair of change-points is the copy number of the corresponding segment. Copy number profiling by microarray and sequencing techniques gives noisy estimates of the true copy number at specific genomic positions, which can be modeled as follows:

Assume a series of n log ratio copy number estimates $\mathbf{x} = \{x_i : i = 1, 2, \cdots, n\}$ ordered by genomic position. The piecewise constant model for the series is

$$x_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where $\boldsymbol{\mu} = \{\mu_i : i = 1, 2, \cdots, n\}$ is a piecewise constant function and $\boldsymbol{\varepsilon} = \{\varepsilon_i : i = 1, 2, \cdots, n\}$ is a sequence of independent and identically distributed errors. Assuming a series of $m + 1$ change-points $\boldsymbol{\tau} = \{\tau_j : j = 0, 1, 2, \cdots, m\}$ where $1 = \tau_0 < \tau_1 < \cdots < \tau_m = n + 1$ delimit m segments with copy number level $\boldsymbol{\theta} = \{\theta_j : j = 1, 2, \cdots, m\}$ such that

$$\mu_i = \theta_j, \quad i \in [\tau_{j-1}, \tau_j), \quad j = 1, 2, \cdots, m. \tag{2}$$

The errors are usually assumed to be Gaussian $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ and supported by experimental data on self-self hybridizations [13], although this assumption is not crucial if the distances between successive τ_j's are large [14]. The DLRS is denoted by σ and estimated by the standard deviation of the error $\boldsymbol{\varepsilon}$.

Segmentation is applied to recover the genomic position of the boundaries and the underlying copy number for chromosomal segments from the noisy copy number estimates. Under the piecewise constant model, the segmentation problem is to find the change-points $\boldsymbol{\tau}$ delimiting the segments and the copy number levels $\boldsymbol{\theta}$ for each segment. If $\boldsymbol{\tau}$ is known, θ_j can be estimated by the mean of the copy number estimates that fall in the j-th segment, that is

$$\hat{\theta_j} = \frac{\sum_{i=\tau_{j-1}}^{\tau_j - 1} x_i}{\tau_j - \tau_{j-1}}, \quad j = 1, 2, \cdots, m. \tag{3}$$

Many segmentation algorithms have been proposed (see [15] and [16] for excellent review), among which the popular circular binary segmentation (CBS) algorithm [11, 12] was found to be one of the most accurate methods. Starting with the whole

chromosome, the CBS algorithm detects change-points delimiting a sub-segment with a different copy number level in the middle of a larger segment, and does it recursively until no more change-points can be found in any of the segments. For any interval $[a, b)$ and $1 \leq a < b \leq n$, let the null hypothesis be that the log ratio copy number estimates $x_a, x_{a+1}, \cdots, x_b$ are independent and identically distributed Gaussian and the alternative be that there is a sub-segment with different mean and same variance. CBS calculates the maximal t-statistic $T = \mathsf{max}_{a \leq i < j \leq b} |T_{ij}|$,

$$T_{ij} = \frac{\bar{Y}_{ij} - \bar{Z}_{ij}}{s_{ij}\sqrt{(j-i)^{-1} + (b-a+i-j+1)^{-1}}}, \tag{4}$$

where $\bar{Y}_{ij} = (x_{i+1} + \cdots + x_j)/(j-i)$, $\bar{Z}_{ij} = (x_a + \cdots + x_i + x_{j+1} + \cdots + x_b)/(b - a + i - j + 1)$, and s_{ij}^2 is the corresponding mean squared error.

Step detection and step-likeness quantification by step-fitting

Ideally, DNA is uniformly retrieved from the chromosomes during sample preparation, that is, the amount of DNA retrieved is proportional to the true copy number of the chromosomal segments present in the cells. In this case, the log ratio copy number estimates contain only abrupt jumps which satisfy the piecewise constant assumption. However, in practice, due to GC-content bias [4], DNA replication timing [5] and other biological phenomena such as differences in chromatin organization during DNA isolation [6] or damages to the DNA caused by formalin fixation [7], DNA is sometimes retrieved differentially across the genome, which adds artefacts in the form of waves to the otherwise piecewise constant signal. These wave artefacts adversely affect detecting change-points which truly delimit the gained and lost chromosomal segments.

Existing segmentation algorithms such as CBS are poorly suited to discriminate the change-points that are boundaries of the CNA segments from those introduced by wave artefacts. When the magnitude of the waves are less than that of the CNA signal, CBS leads to "hyper-segmentation", in which many change-points caused by waves are detected in a single CNA segment. In the worst case where the true copy number signal is buried in the waves with comparable or even greater magnitude, change-points detected by CBS largely depend on the wave artefacts and are no longer boundaries of the CNA segments.

Fortunately, in practice, CNAs induce abrupt jumps which are visualized as steps and wave artefacts only induce gradual changes. As a result, for copy number profiles containing CNAs, reliable samples are usually step-like, with few waves in each CNA segment, whereas unreliable ones contain a lot of waves or the overall shape of the signal does not resemble steps any more. Based on these observations we implemented a fast step-fitting algorithm of time complexity $O(n\log n)$ adapted from the method originally proposed by Kerssemakers et al. (refer to Supplementary Methods 3 in [17]) to capture the step signal which are mostly CNAs regardless of the change-points introduced by wave artefacts, and assess how much the copy number profile is step-like. See Materials and methods section for a detailed explanation about the step-fitting algorithm and the adaptation.

Computer simulation: CBS versus the step-fitting algorithm

To demonstrate the differences between CBS and the step-fitting algorithm in detecting change-points, 3 groups of copy number profiles each containing 200 samples of 10,000 dimensions were generated (Figure 1 and 2). Group A are reliable copy number profiles containing many CNAs; Group B are unreliable copy number profiles in which the piecewise constant CNA signal is buried in waves; Group C are hyper-segmented copy number profiles in which many change-points caused by waves are present in a single copy number segment. See Supplementary Methods for details on how simulations were performed.

Figure 1 shows a set of simulated copy number profiles having the same true CNA segments. A1 to A3 demonstrate the segmentation on the same reliable copy number profile, where the red lines in A1 are the predefined true CNA segments. Red lines in A2 and A3 are copy number segments recovered by CBS and step-fitting respectively, from which it can be concluded that both methods recovered the majority of copy number segments. While CBS performs better in terms of detecting small segments of low signal to noise ratio, step-fitting is capable of capturing the overall step-like signal. B1 to B3 show the same unreliable copy number profile with heavy wave artefacts generated by introducing large auto-correlation to the reliable copy number profile in group A, for which neither CBS nor step-fitting can locate the true CNA segments correctly. In B2, CBS found greater amount of change-points most of which are noise, whereas in B3 step-fitting also fitted noise but only those with more abrupt changes and therefore the number of change-points detected are less than CBS; Nevertheless, change-points detected by both methods mainly depend on noise and are no longer boundaries of the true CNA segments. C1 to C3 show the same hyper-segmented copy number profile created by passing the reliable copy number profile in group A through a median filter to add small waves at the same time preserving the overall step-like structure. While in C2 CBS detected too many change-points within individual true CNA segments, step-fitting significantly outperformed by recovering majority of copy number segments well. Note here the possibility of merging multiple segments detected by CBS using additional methods is not considered, as our main goal is to find a good proxy for the extent of wave artefacts which is represented by the number of change-points that CBS can detect under its standard setting.

The observation from Figure 1 can be generalized to the set of 200 simulated copy number profiles as shown in Figure 2. For reliable copy number profiles in group A, the number of segments recovered by CBS is a very good approximation to the true CNA segments (Spearman correlation coefficient $\rho = 0.89$; Regressing the number of recovered segments on the number of true segments by robust linear regression results in slope $w = 0.93$ and intercept $b = 0.28$), while step-fitting recovered majority of CNA segments well except for those very short segments of low signal to noise ratio ($\rho = 0.41$; $w = 0.67$ and $b = 0.79$). For copy number profiles in group B containing heavy wave artefacts, both methods fitted noise. CBS recovered more segments than the number of true segments ($w = 1.86$ and $b = -2.26$) and although in our simulation there is still correlation ($\rho = 0.70$) as the waves introduced in different samples are of comparable size, the correlation is not expected in real data where the waves are of various size coming from different

sources. Furthermore, in group B, the number of segments found by step-fitting is largely less than CBS and independent of the number of true segments ($\rho = 0.04$; $w = 0.13$ and $b = 19.56$). Last, for hyper-segmented copy number profiles in group C, CBS recovered far more segments than the number of true segments ($\rho = 0.36$; $w = 2.84$ and $b = 222.80$), whereas step-fitting found majority of CNA segments well ($\rho = 0.38$, $w = 0.79$ and $b = 2.21$).

Reliability assessment metrics

In this section we further develop the reliability assessment method on 1522 previously published copy number profiles from different experimental batches and microarray platforms (See Materials and methods section for detail).

Figure 3 shows five typical cases each corresponding to one of the reliability groups that can often be observed among genomic copy number profiles. Apart from those containing many CNAs which can be classified as hyper-segmented, reliable and unreliable with heavy wave artefacts (Figure 3a to 3c), there exist copy number profiles that have no or very few CNAs (e.g. control samples from normal tissues; Figure 3d); but also copy number profiles with extraordinarily large DLRS suggesting poor DNA quality prohibiting the detection of any CNA (Figure 3e). The upper panel of each subgraph in Figure 3a to 3e shows copy number profile of a particular case segmented by CBS, and the lower panel shows the same copy number profile segmented by step-fitting in the optimal iteration when S_{peak}, the maximum value of S generated by the step-fitting algorithm was attained. The corresponding S values throughout iterations are plotted in Figure 3f (See The step-fitting algorithm in Materials and methods section for definition of S and S_{peak}).

In Figure 3a to 3c, the performance of CBS versus step-fitting is consistent with that on simulated data in the previous section: In reliable copy number profiles (Figure 3b), segments recovered by CBS are comparable to those from step-fitting; in unreliable copy number profiles with heavy wave artefacts (Figure 3c) both methods fitted noise in which CBS found more change-points than step-fitting; in hyper-segmented copy number profiles (Figure 3a) CBS detected a great many change-points within individual steps recovered by step-fitting, resulting in a higher number of CBS derived segments. An additional feature associated with reliability can be discovered by looking closer into the three cases in which both the reliable and the hyper-segmented copy number profile have clear step-like structures, while in the unreliable case the boundaries of the CNA segments are much less defined. This property is reflected by the peak value of S: As shown in Figure 3f, the hyper-segmented (Case 1) and the reliable copy number profile (Case 2) both attained relatively high S_{peak} whereas S_{peak} for the unreliable copy number profile (Case 3) is much lower.

In Figure 3d and 3e, no apparent CNA exists or can be detected. The former is unreliable due to large DLRS so that the CNAs, even if are present, cannot be detected properly, whereas the latter is a reliable control sample diploid across the genome. Nevertheless, in both cases step-fitting merely fitted noise and the corresponding S values for Case 4 and 5 in Figure 3f increase throughout iterations yet remain close to 1 and no apparent peak can be observed.

So far we have discussed 4 features related to the reliability of copy number profiles, i.e. S_{peak}, an indicator of how much the copy number profile is step-like;

l, the number of steps detected by step-fitting; v, the number of segments detected by CBS which could be induced by both CNAs and wave artefacts; and σ, which is the DLRS estimated by the standard deviation of the error ε in equation (1).

The full set of reliability assessment metrics is summarized in Table 1, where l and v combined represent the wave density.

CNARA: reliability assessment for copy number profiles

To turn the qualitative metrics into quantitative ones and therefore allow to predict the reliability of a given copy number profile, a support vector machine (SVM) classifier was trained on the 1522 previously published copy number profiles labeled as reliable or unreliable by experts (see Materials and methods section).

The 4 features S_{peak}, l, v and σ were extracted for each of the 1522 samples and were used with the SVM classifier. The *svm* function in the *e1701* R package [18] which is the interface to the C++ implementation *libsvm* [19] was called, where the cost was set to 1 and the radical basis function (RBF) kernel $exp(-0.4|u-v|^2)$ was chosen in which the parameters were set by cross validation. Tenfold cross-validation on the 1522 training samples resulted in a total prediction accuracy of 99.08% with standard deviation 0.0083 (a 3D visualization can be found in Supplementary Figure S5). The extremely high accuracy suggests that the 4 features chosen can represent the copy number profile very well for reliability assessment. The resulted classifier can predict any new copy number profiles as reliable or unreliable in terms of whole-genome CNA evaluation. For fine tuning of the methodology to particular sample compositions or evaluation goals, we have included the option to use user provided training sets (see respective subsection below).

For a copy number profile predicted as reliable, the value of S_{peak} tells if it has many CNAs or not (case 2 and 5 in Table 1). Usually control samples from normal tissues or samples without many CNAs have $S_{peak} \leq thr_1$, while tumor samples containing CNAs that we are interested in have $S_{peak} > thr_1$, where thr_1 is a threshold with default value 1.5.

Copy number profiles predicted as unreliable can be further divided into 3 subcategories (case 1, 3 and 4 in Table 1), by carrying out the following procedure: Denote samples in the training set by $(S_{peak}^{(i)},\ l^{(i)},\ v^{(i)},\ \sigma^{(i)})$ where $i = 1,\ 2,\ \cdots,\ 1522$. Given any unreliable sample $b = (S_{peak},\ l,\ v,\ \sigma)$, create two dummy samples $d_1 = (S_{peak},\ l,\ min(v^{(i)}),\ \sigma)$ and $d_2 = (S_{peak},\ l,\ v,\ min(\sigma^{(i)}))$ by substituting v or σ by the minimum of the training set correspondingly, and predict the reliability of d_1 and d_2. If d_1 is reliable, b contains wave artefacts and therefore can be subcategorized as either Case 1 or 3; if d_2 is reliable, b has large DLRS and thus belongs to Case 4; otherwise b suffers from both wave artefacts and large DLRS and can be either Case 1 or 3. As hyper-segmented samples in Case 1 may still be of interest for certain tasks such as looking for CNA regions but not counting the number or keeping track of the size of copy number gains or losses, a more stringent threshold thr_2 greater than thr_1 for S_{peak} could be set, for example, to accept those hyper-segmented samples having $S_{peak} > thr_2$ with caution where the default value of thr_2 is 2.5.

Discussion

A comparison between CNARA and medASP

The performance of CNARA and medASP [7] was compared and shown in Figure 4. The 1522 samples were randomly split into training and validation sets at the proportion of 50:50% and the SVM classifier of CNARA was trained on the training set where the parameters for the cost and the RBF kernel were set the same as in the previous subsection. The probability of being predicted as unreliable for each sample in the validation set was computed by the *libsvm* implementation [19] and the median of ASP was computed for the same sample in the validation set. The receiver operating characteristic (ROC) curves for the two values were then generated against the true class labels for the validation set.

As shown in Figure 4, the area under the ROC curve (AUC) for medASP is 0.7372, which suggests that the assumptions made in the medASP study [7], i.e. simulating copy number profiles with a gain and a loss region of fixed size, does not comform well with feature distribution in real-world data. By contrast, the AUC of CNARA is 0.9994 which is consistent with the extremely high prediction accuracy of the SVM classifier stated in the previous subsection. The value approaching 1 once again implies that the 4 features chosen summarize the copy number profile very well in terms of its reliability.

Custom training set

Apart from the training set consists of absolutely reliable and unreliable copy number profiles, CNARA is flexible to take in additional average quality samples labeled by experts to the training set to accomodate to their own need. For example, the boundaries of the CNA segments for formalin-fixed paraffin embedded (FFPE) samples are known to be less well defined than fresh-frozen samples in general [20, 7]. Therefore, when working with a large set of FFPE samples, one may wish to keep those slightly contaminated samples and reject others that are highly contaminated. To achieve this, people can create their own training set by including several slightly contaminated samples as reliable and highly contaminated as unreliable according to expert knowledge to the original training set (See Supplementary Methods for more details on building custom training set).

Conclusions

During our efforts on the curation of human cancer copy number data from genomic screening experiments [21, 22], we have frequently encountered data sets challenging the state-of-the-art quality assessment procedures. While several methods have been proposed to improve the readout of genomic copy number profiling both through improvements of experimental design [6] as well as application of advanced bioinformatic methods [3, 4, 8, 9], many data sets still contain artefacts that can not be removed sufficiently by these existing methods. As a result, when working with tens of thousands of genomic copy number profiles derived from a multitude of platforms and different pre-processing methods, a robust method capable of identifying the low quality data sets based on extractable features is needed.

Previously, some downstream quality control measures such as genotyping call rate [23] and derivative log-ratio spread (DLRS) [10] have been reported for microarrays

to describe the noise induced during sample hybridization onto the arrays. However, existing evaluation methods are limited in controlling for inherent noise due to differential DNA retrieval across the genome during sample preparation, which is a severe problem in the generation of genomic copy number profiles, and therefore reliability assessment for copy number profiles remained an open problem. Zhang and Zhang previously did an explorative study [7] where they proposed medASP as a discriminative metric for evaluating the quality of copy number profiles and achieved good accuracy on simulated data. However, it has been shown in the previous section that medASP is not an adequate measure for real copy number data which contains different amount of copy number gains and losses of various sizes at different genomic positions.

In this paper, we proposed a method for assessing the reliability of DNA copy number profiles. We showed five typical cases each corresponding to one of the reliability groups that can often be observed in practice and discussed in detail about the 4 features which can represent the copy number profiles well in terms of reliability, namely the number of steps detected by step-fitting, an indicator of how much the copy number data is step-like, the number of segments induced by both CNAs and wave artefacts detected by CBS and the mean of DLRS within segments. To obtain the first two features we proposed a fast step-fitting algorithm of time complexity $O(n\log n)$ which is scalable to high-throughput copy number data. Taking these 4 features as input, an SVM classifer was trained on 1522 samples labeled as reliable or unreliable according to expert knowledge. By tenfold cross-validation on the whole dataset, the resulting classifier achieved a total accuracy of 99.08% with standard deviation 0.0083. Predicted samples can be further subcategorized into the five reliablility classes, each having a biological or experimental correspondence. To the best of our knowledge, this is the first application developed for real-world data filling the gap of controlling the quality problem regarding differential DNA retrieval for copy number profiles, which is non-overlapping with and complementary to the objectives of any state-of-the-art quality control measures.

Our method can be applied to log ratio copy number data before or after normalization. Applied before normalization, it helps to judge if the log ratio copy number data needs normalization and noise correction; applied to the normalized data, it assesses the utility of the data in downstream analysis. Nonetheless, since it evaluates an aspect of the quality different from any existing quality control measures, we suggest our reliability assessment method to be the last step after any possible downstream quality control and bias correction methods, to decide if the normalized sample can be finally included in data mining experiments for biomedical knowledge generation. We have to emphasize that, while our method can deliver information about the quality of the signal derived from whole genome copy number screening experiments, by itself it does not address problems arising from the possible clonal heterogeneity, e.g. in biosamples derived from cancer tissues. Also, the impact of reliability assessment will depend on the intended downstream analyses; for instance, the reliability of whole-genome CNA profiles, as determined by our method, may be of less concern when using statistical CNA peak finding tools like GISTIC [24].

Thanks to the competing efforts on estimating genomic copy number from exome and whole-genome sequencing [25] especially recent development of CopywriteR [26]

which is capable of extracting uniformly distributed copy number information from sequencing data, we are optimistic that in the near future piecewise constant signal for copy number estimates could be computed reliably with comparable acurracy to that of microarray platforms. Due to the universality of our method in dealing with samples from different platforms and the flexibility in taking new training samples, at that time our reliability assessment method can also be easily adapted to those copy number profiles generated by sequencing platforms.

Materials and methods

The copy number profile dataset

1522 previously published copy number profiles (Supplementary Table S1) were used in our study. The samples were obtained from arrayMap [21, 22] with the original data being having been retrieved from NCBI's GEO repository [27], with preprocessing and noise correction performed through the standard arrayMap data processing pipeline [21]. Probe-level plots with added segmentation markers were visually inspected and selected by experts with respect to the empirically assessed copy-number calling reliability, and with the goal of a balanced representation of reliable and unreliable copy number profiles as well as a sufficient coverage of cases from different reliability groups. This selection resulted in 804 absolutely reliable copy number profiles (cf. cases 2 and 5 in Table 1) and 718 absolutely unreliable copy number profiles including hyper-segmented ones (cf. cases 1, 3 and 4 in Table 1). See Supplementary Methods for data preprocessing and Table S1 for a complete list of the arrays being analyzed and their reliability labels. The data is available at arraymap.org [21, 22].

The step-fitting algorithm

To recover the set of change-points $\boldsymbol{\tau} = \{\tau_j : j = 0, 1, 2, \cdots, m\}$ delimiting the CNA segments from the log ratio copy number estimates $\mathbf{x} = \{x_i : i = 1, 2, \cdots, n\}$, the algorithm updates $\boldsymbol{\tau}$ iteratively, starting from $\boldsymbol{\tau}^{(0)} = \{1, n+1\}$, in each iteration k introduces an additional change-point $\tau^{(k)}$ to $\boldsymbol{\tau}^{(k-1)}$, called the best-fit, which minimizes the cost function

$$H = \sum_{i=1}^{n} (x_i - \mu_i)^2 \tag{5}$$

by scanning through all possible locations $i = 2, \cdots, n$, $i \notin \boldsymbol{\tau}^{(k-1)}$ and adding i to $\boldsymbol{\tau}^{(k-1)}$ as the temporary change-point set $\boldsymbol{\tau}^{(k)}_{temp}$ such that $\boldsymbol{\tau}^{(k)}_{temp} = \text{sort}(\boldsymbol{\tau}^{(k-1)} \cup i)$, in which the elements are ordered from the smallest to the largest; μ_i in equation (5) is computed from equation (2) and (3) where $\{\tau_{j-1}, \tau_j\} \subseteq \boldsymbol{\tau}^{(k)}_{temp}$, $j = 1, 2, \cdots, k+1$. The location i that minimizes equation (5) is therefore $\tau^{(k)}$, and $\boldsymbol{\tau}$ in the kth iteration is updated as $\boldsymbol{\tau}^{(k)} = \text{sort}(\boldsymbol{\tau}^{(k-1)} \cup \tau^{(k)})$.

Next, in between each pair of τ_{j-1} and τ_j in $\boldsymbol{\tau}^{(k)}$ where $j = 1, 2, \cdots, k+1$, find the change-point c_j which is the best-fit for x_i, $i \in [\tau_{j-1}, \tau_j)$ such that it minimizes

$$F_j = \sum_{i=\tau_{j-1}}^{\tau_j - 1} (x_i - \psi_i)^2, \tag{6}$$

where $c_j \in (\tau_{j-1}, \tau_j)$ and

$$\psi_i = \begin{cases} \sum_{i=\tau_{j-1}}^{c_j-1} x_i/(c_j - \tau_{j-1}) & \text{for } i \in [\tau_{j-1}, c_j) \\ \sum_{i=c_j}^{\tau_j-1} x_i/(\tau_j - c_j) & \text{for } i \in [c_j, \tau_j) \end{cases} \tag{7}$$

The set of change-points c_j's plus the boundary denoted by $\boldsymbol{c}^{(k)} = \{c_j : j = 1, 2, \cdots, k+1\} \cup \{c_0 = 1,\ \ c_{k+2} = n+1\}$ is called the counter-fit change-points for iteration k. The cost function Q for the counter-fit is defined as

$$Q = \sum_{i=1}^{n} (x_i - \nu_i)^2. \tag{8}$$

where

$$\nu_i = \frac{\sum_{i=c_{j-1}}^{c_j-1} x_i}{c_j - c_{j-1}}, \quad i \in [c_{j-1}, c_j), \quad j = 1, 2, \cdots, k+2. \tag{9}$$

The algorithm proceeds iteratively adding the best-fit change-point $\tau^{(k)}$ each time to the change-point set $\boldsymbol{\tau}^{(k-1)}$ in the previous iteration and finding the set of counter-fit change-points $\boldsymbol{c}^{(k)}$ correspondingly, until the number of iterations reaches a predefined threshold K. This results in a set of best-fit change-points and a set of counter-fit change-points located in between one another.

The step-fitting algorithm stated above (refer to Supplementary Methods 3 in [17] for more information) has time complexity of $\sim 2nK$ (for definition of the tilde notation see [28]) where n is the dimension of the copy number estimates $\mathbf{x}$ and K is the total number of predefined iterations usually greater than 100. An important observation which helps to improve the computational efficiency is that, of all the counter-fit change-points in $\boldsymbol{c}^{(k)}$ in the kth iteration, the most prominent change-point c_j if added to $\boldsymbol{\tau}^{(k)}$ decreasing the cost function H the most is always the best-fit change-point $\tau^{(k+1)}$ to be included in the next iteration, specifically,

$$\tau^{(k+1)} = \underset{c_j \in \boldsymbol{c}^{(k)}}{\arg\max}\, d_j \tag{10}$$

and

$$d_j = \sum_{i=\tau_{j-1}}^{\tau_j-1} (x_i - \mu_i)^2 - (x_i - \psi_i)^2, \tag{11}$$

where $i \in [\tau_{j-1}, \tau_j)$, $\{\tau_{j-1}, \tau_j\} \subseteq \boldsymbol{\tau}^{(k)}$, $c_j \in (\tau_{j-1}, \tau_j)$ and $j = 1, 2, \cdots, k+1$; μ_i and ψ_i are computed as in equation (2), (3) and (7) respectively. Here we keep track of the set of d_j's for each corresponding c_j as the difference set $\boldsymbol{d}^{(k)}$. Furthermore, after the best-fit change-point $\tau^{(k+1)}$ being included in $\boldsymbol{\tau}^{(k+1)}$, to update the corresponding counter-fit set $\boldsymbol{c}^{(k+1)}$ and difference set $\boldsymbol{d}^{(k+1)}$ we only need to exclude $\tau^{(k+1)} = c_t$ from $\boldsymbol{c}^{(k)}$ and the corresponding d_t from $\boldsymbol{d}^{(k)}$ first, and then scan the region delimited by the two best-fit break-points directly next to $\tau^{(k+1)}$ in $\boldsymbol{\tau}^{(k+1)}$

for two additional counter-fit break-points and the corresponding two differences. To be specific, given $\tau_r = \tau^{(k+1)}$, $\{\tau_{r-1}, \tau_r, \tau_{r+1}\} \subseteq \boldsymbol{\tau}^{(k+1)}$, find $c_r \in [\tau_{r-1}, \tau_r)$ and $c_{r+1} \in [\tau_r, \tau_{r+1})$ as computed in equation (6) and (7), and compute d_r and d_{r+1} as in equation (11); then update $\boldsymbol{c}^{(k+1)} = (\boldsymbol{c}^{(k)} \setminus \tau^{(k+1)}) \cup c_r \cup c_{r+1}$ and $\boldsymbol{d}^{(k+1)} = (\boldsymbol{d}^{(k)} \setminus d_t) \cup d_r \cup d_{r+1}$. In this way, the time complexity is reduced to $O(n\log n)$.

To estimate the number of steps in the copy number profile and assess how much the data is step-like, the same model selection criterion is adopted as in Kerssemakers' method, where the step-indicator S is introduced and defined as

$$S = \frac{Q}{H}. \tag{12}$$

For reliable copy number profiles or hyper-segmented profiles containing many steps which are mostly CNAs, in early iterations change-points in $\boldsymbol{\tau}^{(k)}$ and $\boldsymbol{c}^{(k)}$ both locate significant steps such that Q is close to H and therefore S is close to 1. S increases until it peaks in the optimal iteration when change-points in $\boldsymbol{\tau}^{(k)}$ cover all significant steps and all change-points in $\boldsymbol{c}^{(k)}$ merely fit noise so that Q and H differ the most. After that S decreases as change-points in both $\boldsymbol{\tau}^{(k)}$ and $\boldsymbol{c}^{(k)}$ begin to fit noise and Q and H become closer again. As a result, the number of iterations it takes for S to reach the peak, denoted by l, is an estimation of the number of steps in the copy number profile.

In Kerssemakers' literature they also mentioned the peak value of S denoted by S_{peak} approximates quadratic of signal to noise ratio given by $S_{peak} \approx 1 + \frac{\Delta^2}{4\sigma^2}$, where Δ is the mean of the absolute difference of μ_i's around each best-fit change-point τ_j weighted by $\sqrt{\tau_{j+1} - \tau_{j-1}}$, where $\tau_j \in \boldsymbol{\tau}^{(l)}$, $j = 1, 2, \cdots, l$, and σ is the standard deviation of $\boldsymbol{\varepsilon}$ in equation (1). For unreliable copy number profiles of which the steps are drowned by wave artefacts of large magnitude, S_{peak} is significantly smaller than that of reliable copy number profiles containing comparable amount of CNAs. For those profiles containing few significant CNAs or control samples which are diploid across genome, S is close to 1 throughout iterations and no apparent peak can be observed so that S_{peak} is close to 1. Therefore S_{peak} is an indicator of how much the copy number profile is step-like.

The adapted step-fitting algorithm is summarized in Algorithm 1.

Algorithm 1 Adapted step-fitting algorithm

Input: $\mathbf{x}$
Initialize: $\boldsymbol{\tau}^{(0)} = \{1, n+1\}$, $\boldsymbol{c}^{(0)} = \{1, c_1, n+1\}$, $\boldsymbol{d}^{(0)} = \{d_1\}$ and $S^{(0)} = 1$.
For $k = 1, 2, \cdots, K$
1: Find $\tau^{(k)} = \arg\max_{c_j \in \boldsymbol{c}^{(k-1)}} d_j$ where $d_j \in \boldsymbol{d}^{(k-1)}$, $j = 1, 2, \cdots, k$
2: Set $\boldsymbol{\tau}^{(k)} = \mathsf{sort}(\boldsymbol{\tau}^{(k-1)} \cup \tau^{(k)})$
3: Given $\tau_r = \tau^{(k)} = c_t \in \boldsymbol{c}^{(k-1)}$, $\{\tau_{r-1}, \tau_r, \tau_{r+1}\} \subseteq \boldsymbol{\tau}^{(k)}$, find $c_r \in [\tau_{r-1}, \tau_r)$, $c_{r+1} \in [\tau_r, \tau_{r+1})$, d_r and d_{r+1}
4: Update $\boldsymbol{c}^{(k)} = (\boldsymbol{c}^{(k-1)} \setminus \tau^{(k)}) \cup c_r \cup c_{r+1}$
5: Update $\boldsymbol{d}^{(k)} = (\boldsymbol{d}^{(k-1)} \setminus d_t) \cup d_r \cup d_{r+1}$
6: Set $S^{(k)} = Q/H$
Output: $l = \arg\max_k \{S^{(k)}, k = 1, 2, \cdots, K\}$, $S_{peak} = S^{(l)}$, $\boldsymbol{\tau}^{(l)}$, $\boldsymbol{c}^{(l)}$

Software

The CNARA software and a tutorial is available at https://github.com/baudisgroup/CNARA.

Abbreviations

CNA/CNV: copy number alterations/variations; DLRS: derivative log ratio spread; ASP: autocorrelation scanning profile; medASP: median of ASP; CBS: circular binary segmentation; SVM: support vector machine; RBF: radical basis function; ROC: receiver operating characteristic; AUC: area under the ROC curve; FFPE: formalin-fixed paraffin embedded; CNARA: reliability assessment for copy number profiles.

Ethics statement

The human array datasets utilized in this study represent publicly available datasets that were derived from the Gene Expression Onmibus (GEO [27]) resource at the National Center for Biotechnology Information (NCBI), with the possibly required informed consent lying with the orginal data submitters. For additional samples used for evaluation purposes (data not shown), informed consent was acquired from the patients with approval of the study design through the ethics committee of the University of Medicine and Pharmacology Timisoara.

Availability of supporting data

The GEO accession numbers of the arrays used in the study are provided in the supplementary file and can be used to access GEO datasets as well as the corresponding processed data through the arrayMap resource. The CNARA software and a tutorial is available at https://github.com/baudisgroup/CNARA.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

NA conceived the study and developed computational methods; HC collected and labeled the data; CS provided additional test samples; NA and MB wrote the paper; All authors contributed to the biological insight, read and approved the final manuscript.

Acknowledgements

The authors want to thank Henrik Bengtsson, Reinhard Furrer and Mark Robinson for helpful discussions. This study was supported by a grant from the Swiss Federation through the Swiss Contribution to the enlarged European Union.

Author details

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. [2]Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, 610064 Chengdu, Sichuan, China. [3]Department of Dermatology, "Victor Babeş" University of Medicine and Pharmacy, Timisoara, Romania.

References

1. Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., Pinkel, D.: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science **258**(5083), 818–821 (1992)
2. du Manoir, S., Speicher, M.R., Joos, S., Schröck, E., Popp, S., Döhner, H., Kovacs, G., Robert-Nicoud, M., Lichter, P., Cremer, T.: Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Human genetics **90**(6), 590–610 (1993)
3. van de Wiel, M.A., Brosens, R., Eilers, P.H., Kumps, C., Meijer, G.A., Menten, B., Sistermans, E., Speleman, F., Timmerman, M.E., Ylstra, B.: Smoothing waves in array CGH tumor profiles. Bioinformatics **25**(9), 1099–1104 (2009)
4. Redon, R., Carter, N.P.: Comparative genomic hybridization: microarray design and data interpretation. DNA Microarrays for Biomedical Research: Methods and Protocols, 37–49 (2009)
5. Koren, A., Handsaker, R.E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., McCarroll, S.A.: Genetic variation in human DNA replication timing. Cell **159**(5), 1015–1026 (2014)
6. van Heesch, S., Mokry, M., Boskova, V., Junker, W., Mehon, R., Toonen, P., de Bruijn, E., Shull, J.D., Aitman, T.J., Cuppen, E., *et al.*: Systematic biases in DNA copy number originate from isolation procedures. Genome biology **14**(4), 33 (2013)
7. Zhang, L., Zhang, L.: Use of autocorrelation scanning in DNA copy number analysis. Bioinformatics **29**(21), 2678–2682 (2013)
8. Marioni, J.C., Thorne, N.P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T.D., Stranger, B.E., Lynch, A.G., Dermitzakis, E.T., *et al.*: Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol **8**(10), 228 (2007)
9. Risso, D., Schwartz, K., Sherlock, G., Dudoit, S.: GC-content normalization for RNA-Seq data. BMC bioinformatics **12**(1), 480 (2011)
10. Largo, C., Saéz, B., Alvarez, S., Suela, J., Ferreira, B., Blesa, D., Prosper, F., Calasanz, M.J., Cigudosa, J.C.: Multiple myeloma primary cells show a highly rearranged unbalanced genome with amplifications and homozygous deletions irrespective of the presence of immunoglobulin-related chromosome translocations. Haematologica **92**(6), 795–802 (2007)
11. Olshen, A.B., Venkatraman, E., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics **5**(4), 557–572 (2004)

12. Venkatraman, E., Olshen, A.B.: A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics **23**(6), 657–663 (2007)
13. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G., Jain, A.N.: Hidden Markov models approach to the analysis of array CGH data. Journal of multivariate analysis **90**(1), 132–153 (2004)
14. Zhang, N.R.: DNA copy number profiling in normal and tumor genomes. In: Jianfeng Feng, F.S. Wenjiang Fu (ed.) Frontiers in Computational and Systems Biology, pp. 259–281. Springer, ??? (2010)
15. Lai, W.R., Johnson, M.D., Kucherlapati, R., Park, P.J.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics **21**(19), 3763–3770 (2005)
16. Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics **21**(22), 4084–4091 (2005)
17. Kerssemakers, J.W., Munteanu, E.L., Laan, L., Noetzel, T.L., Janson, M.E., Dogterom, M.: Assembly dynamics of microtubules at molecular resolution. Nature **442**(7103), 709–712 (2006)
18. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C., Meyer, M.D.: Package 'e1071'. Repository CRAN, February (2014)
19. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27–12727 (2011). Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
20. Mc Sherry, E.A., Mc Goldrick, A., Kay, E.W., Hopkins, A.M., Gallagher, W.M., Dervan, P.A.: Formalin-fixed paraffin-embedded clinical tissues show spurious copy number changes in array-CGH profiles. Clinical genetics **72**(5), 441–447 (2007)
21. Cai, H., Kumar, N., Baudis, M.: arrayMap: a reference resource for genomic copy number imbalances in human malignancies. PLoS One **7**(5), 36944 (2012)
22. Cai, H., Gupta, S., Rath, P., Ai, N., Baudis, M.: arrayMap 2014: an updated cancer genome resource. Nucleic acids research, 1123 (2014)
23. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., *et al.*: Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic epidemiology **34**(6), 591–602 (2010)
24. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J., Huang, J., Alexander, S., Du, J., Kau, T., Thomas, R., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R., Demichelis, F., Hatton, C., Rubin, M., Garraway, L., Nelson, S., Liau, L., Mischel, P., Cloughesy, T., Meyerson, M., Golub, T., Lander, E., Mellinghoff, I., Sellers, W.: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A **104**(50), 20007–20012 (2007)
25. Zhao, M., Wang, Q., Wang, Q., Jia, P., Zhao, Z.: Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC bioinformatics **14**(Suppl 11), 1 (2013)
26. Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., Nevedomskaya, E., Xu, G., de Ruiter, J., Lolkema, M.P., *et al.*: CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol **16**(1), 49 (2015)
27. Edgar, R., Domrachev, M., AE, L.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)
28. Sedgewick, R., Wayne, K.: Algorithms, 4th edition., pp. 1–955. Addison-Wesley (2011)
29. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: NCBI GEO: archive for functional genomics data sets–update. Nucleic acids research **41**(D1), 991–995 (2013)
30. Northcott, P.A., Nakahara, Y., Wu, X., Feuk, L., Ellison, D.W., Croul, S., Mack, S., Kongkham, P.N., Peacock, J., Dubuc, A., *et al.*: Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. Nature genetics **41**(4), 465–472 (2009)
31. Chen, R., Nishimura, M.C., Bumbaca, S.M., Kharbanda, S., Forrest, W.F., Kasman, I.M., Greve, J.M., Soriano, R.H., Gilmour, L.L., Rivers, C.S., *et al.*: A hierarchy of self-renewing tumor-initiating cell types in glioblastoma. Cancer cell **17**(4), 362–375 (2010)
32. Liu, L., Greger, J., Shi, H., Liu, Y., Greshock, J., Annan, R., Halsey, W., Sathe, G.M., Martin, A.-M., Gilmer, T.M.: Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. Cancer research **69**(17), 6871–6878 (2009)
33. Haverty, P.M., Fridlyand, J., Li, L., Getz, G., Beroukhim, R., Lohr, S., Wu, T.D., Cavet, G., Zhang, Z., Chant, J.: High-resolution genomic and expression analyses of copy number alterations in breast tumors. Genes, Chromosomes and Cancer **47**(6), 530–542 (2008)

Figures

Tables

Additional Files

Additional file 1 — Supplementary Methods
This document describes the computer simulation procedure for the 3 groups of copy number profiles in Figure 1 and 2, the preprocessing procedure for the 1522 copy number profile dataset and the supporting Figures S1-S5, and the procedure of building custom training set.

Additional file 2 — Supplementary Table S1
A complete list of the 1522 copy number profiles, including GEO accession number and reliability label.

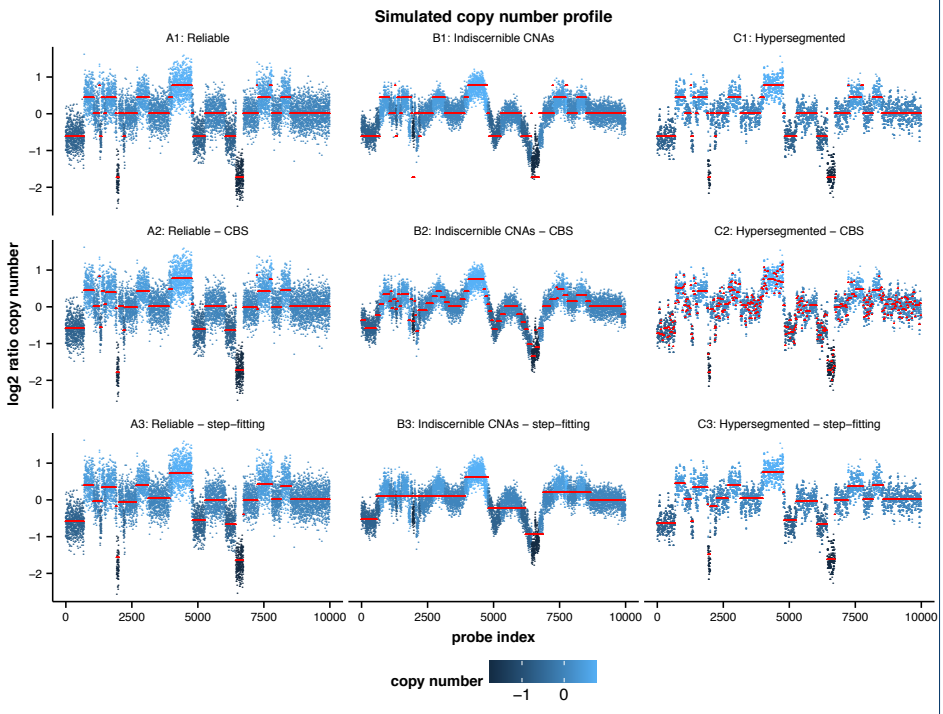


Figure 1: **An example of simulated copy number profiles.** Samples of 10,000 dimensions ($n = 10,000$) were generated for 3 reliability groups, i.e. A: reliable copy number profiles containing many CNAs; B: unreliable copy number profiles having indiscernible CNAs due to wave artefacts; C: hyper-segmented copy number profiles. In group A, both CBS and step-fitting recovered majority of copy number segments well; In group B, both methods fitted noise and CBS detected more change-points than step-fitting; In group C, CBS detected too many change-points whereas step-fitting recovered majority of segments well.

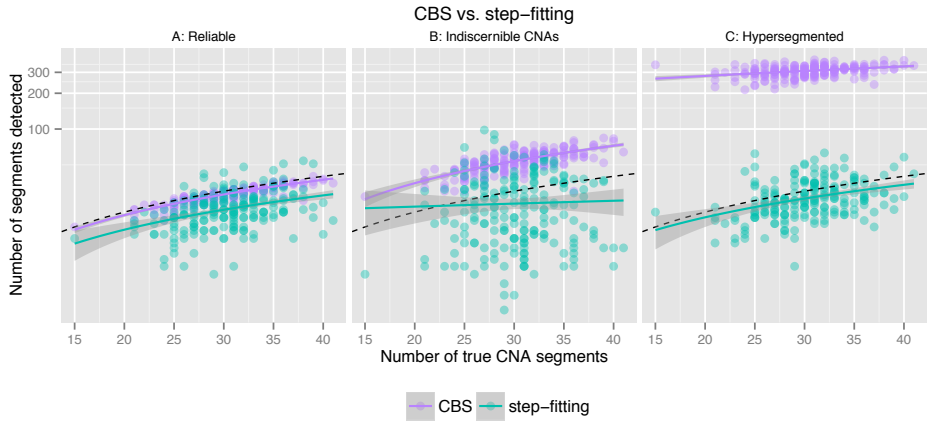
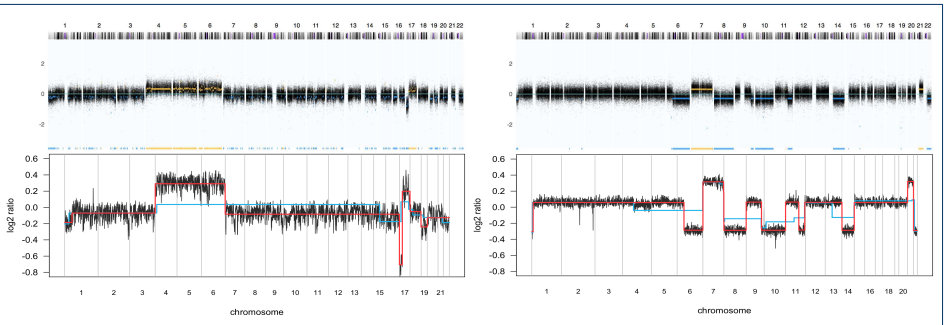


Figure 2: **CBS versus step-fitting for 200 set of simulated copy number profiles.** Scatter plot of the number of segments detected by CBS (purple) and step-fitting (green), plotted against the number of true CNA segments. In group A, the number of segments recovered by CBS approximates true CNA segments very well; the number of segments recovered by CBS is close to that of step-fitting. In group B, both methods fitted noise in which CBS found more change-points than step-fitting in general. In group C, CBS recovered much more segments than the number of true segments, whereas step-fitting found majority of CNA segments well.

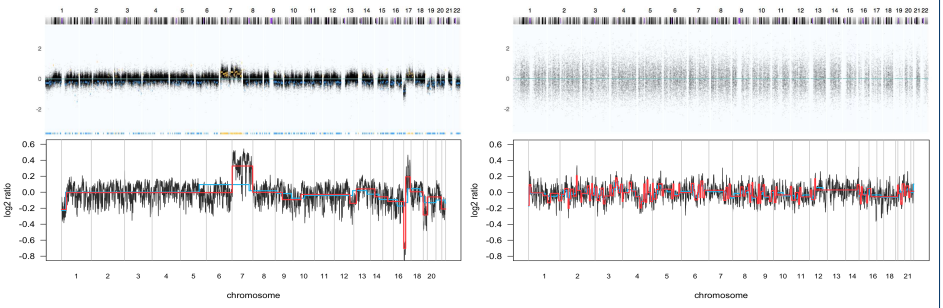
Case No.	S_{peak}	Wave density (l and v)	σ	Assessment
1	high	high	high	hyper-segmented, discernible CNAs with some waves
	high	high	low	
2	high	low	high	reliable, discernible CNAs with few waves
	high	low	low	
3	low	high	high	unreliable, indiscernible CNAs with heavy waves
	low	high	low	
4	low	low	high	unreliable, large DLRS
5	low	low	low	reliable, control sample or without many CNAs

Table 1: Reliability assessment metrics for copy number profiles. Eight qualitative combinations lead to five reliability cases. S_{peak} quantifies how much the copy number data is step-like; Wave density depends on two features l and v, where l is the number of steps recovered by step-fitting in the copy number profile and v is the number of change-points detected by CBS which could be induced by both CNAs and wave artefacts; And σ is the mean of DLRS within segments.



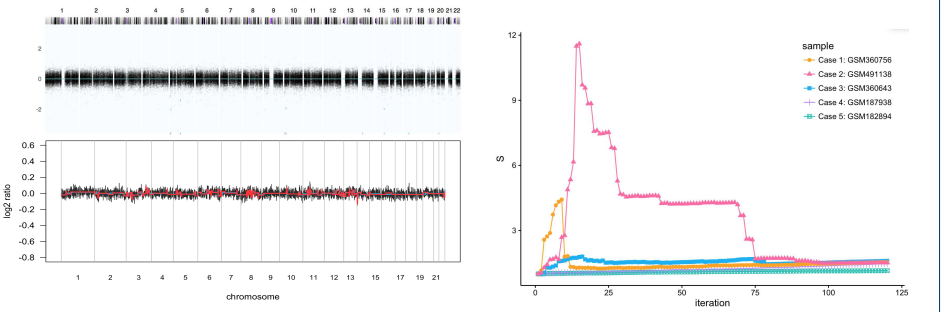
(a) Case 1: hyper-segmented, discernible CNAs with some waves

(b) Case 2: reliable, discernible CNAs with few waves



(c) Case 3: unreliable, indiscernible CNAs with heavy waves

(d) Case 4: unreliable, large DLRS, undetectable CNAs



(e) Case 5: reliable, control sample or without many CNAs

(f) S-plot

Figure 3: **Five example specimen of copy number profiles for each of the reliability groups.** In each subgraph 3a to 3e, the upper panel shows the copy number profile segmented by the CBS algorithm, and the lower panel displays the same copy number profile segmented by step-fitting in the optimal iteration when S_{peak} was attained, where the red line is the fit and the blue line is the counter-fit. In Figure 3f, S values are shown for the same five copy number profiles. For each curve the S-values for 120 iterations are shown. The GEO accession numbers [29] for the five cases are: Case 1, GSM360756 [30]; Case 2, GSM491138 [31]; Case 3, GSM360643 [30]; Case 4, GSM187938 [32]; and Case 5, GSM182894 [33].

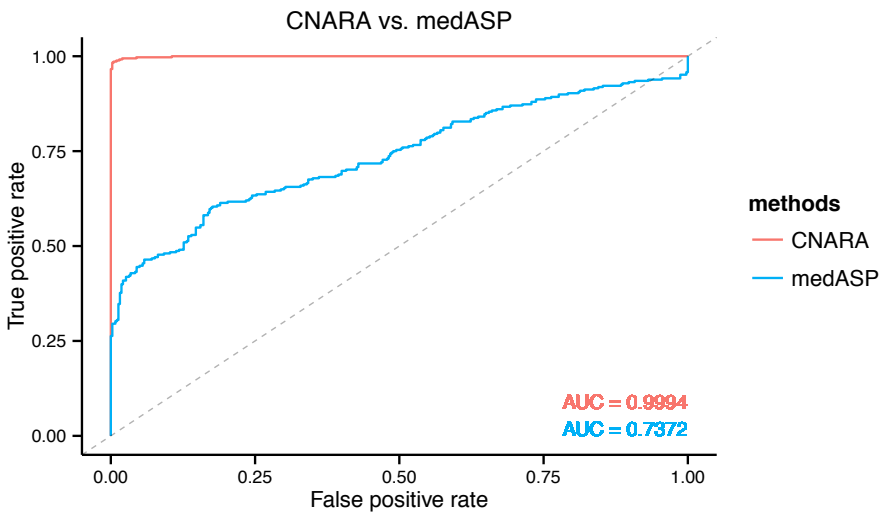


Figure 4: **CNARA versus medASP.** ROC plots of CNARA and medASP on the validation set. The 1522 samples were split into training and validation sets (50:50%) at random, and the SVM classifier of CNARA was trained on the training set. The AUC of CNARA is 0.9994, compared to the AUC of medASP (0.7372).

Supplementary
CNARA: reliability assessment for genomic copy number profiles

Ni Ai [1,*], Haoyang Cai [2], Caius Solovan [3] and Michael Baudis [1,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Switzerland
[2]Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu 610064, Sichuan, China
[3]Department of Dermatology, University of Medicine and Pharmacy "Victor Babes", Timisoara, Romania

[*]to whom correspondence should be addressed

Supplementary Methods

Computer simulation

In the simulation study comparing CBS with the step-fitting algorithm on change-point detection, 3 groups of copy number profiles each containing 200 samples of 10,000 dimensions were generated, in which Group A are reliable copy number profiles with well-defined CNAs; Group B are unreliable copy number profiles contaminated by heavy wave artefacts; Group C are hyper-segmented copy number profiles in which many change-points induced by wave artefacts are present in a single copy number segment. To make the simulation reflect the real data, each of the 200 samples for the 3 groups was generated as follows:

To take into account normal tissue contamination, assume each sample is composed of 30% of normal diploid cells and 70% of cancer cells containing CNAs of true copy number value $CN = (0, 1, 2, 3, 4, 5)$ ranging from full deletion to 5 copies for a particular segment. As a result, the log_2 copy number value each sample can attain is $CNValue = log_2((0.7 * CN + 0.3 * 2)/2)$.

generating true CNA segments

- Sample 50 boundaries from 9,998 positions (10,000 excluding the start and the end position).
- For each segment delimited by two successive boundaries, sample the corresponding copy number value $CNValue$ with probability weight $w = (0.02, 0.15, 0.61, 0.15, 0.05, 0.02)$, where the weight for diploid status is the largest, and decreases as the true copy number value CN gets further from 2.
- If two consecutive segments have the same $CNValue$, merge them into a single segment.

generating a sample for each of the 3 groups

Based on the true CNA segments generated,

- Group A (reliable): Add i.i.d. Gaussian noise $\varepsilon \sim N(0, 0.09)$ to the true CNA segments to obtain x.

- Group B (indiscernible CNAs): Introduce large autocorrelation to x to obtain x_{wave} by adopting a similar strategy as in Zhang and Zhang [1].

$$
\begin{aligned}
&n = 10,000,\ \beta = 0.7,\ lag = 150;\\
&y = (x[lag+1],\ x[lag+2],\ \cdots,\ x[n],\ x[1],\ x[2],\ \cdots,\ x[lag]);\\
&z = (x[n-lag+1],\ x[n-lag+2],\ \cdots,\ x[n],\ x[1],\ x[2],\ \cdots,\ x[n-lag]);\\
&x_{wave} = x*(1-\beta) + (y+z)*\beta/2.
\end{aligned}
$$

- Group C (hyper segmented): Add i.i.d. Gaussian noise $\varepsilon \sim N(0, 0.36)$ to the true CNA segments and pass into a median filter of window size 10 to obtain $x_{hyperseg}$.

Data preprocessing of the copy number profile dataset

The data analyzed in the study consists of 1522 previously published copy number profiles retrieved from arrayMap [2, 3], for which preprocessing and noise correction was done by the project's data pipeline (as shown in Figure 5 of [2]) and the raw data is available at the Gene Expression Omnibus (GEO) [4] website. The total number of the copy number profiles were unknown beforehand. Samples were visually inspected and picked by experts, so that each sample was classified as "reliable" or "unreliable", while the amount of reliable and unreliable copy number profiles was balanced and cases from different reliability groups were well represented. This empirical selection process resulted in 804 äbsolutely reliablecopy number profiles (see ee.g. case 2 and 5 in Table 1 of the main article) and 718 äbsolutely unreliablecopy number profiles including hyper-segmented ones (cases 1, 3 and 4 in Table 1 of the main article). A complete list of the samples and the related information including platforms and reliability labels are given in Supplementary Table S1. Data input to the CBS and the step-fitting algorithm were then processed accordingly.

We observed that CBS tends to detect significantly more breakpoints for platforms of higher resolution such as the Affymetrix Genome-Wide Human SNP 6.0 Array (GEO GPL6801, purple, Figure S1). Thus to accommodate to platforms of different resolution, the first data point in every k_1 data points was kept where k_1 is the rounded value of the total number of data points in a sample divided by $100,000$. In this way, at most $100,000$ data points were kept for each sample and then fed into the CBS algorithm. As shown in Figure S1, the range of the number of change-points detected by CBS for down-sampled data is more consistent among platforms of different resolutions than the original data without down-sampling.

The step-fitting algorithm performed better on samples having greater signal-to-noise ratio (Figure S2). To reduce the noise, the copy number data was smoothed by a median filter of window size 100. To lessen the autocorrelation effect of the smoothed data without losing the entirety of the piecewise constant structure in the data, the first data point in every k_2 data points was kept for the smoothed data where k_2 is the rounded value of the total number of data points in a sample divided by $10,000$ (Figure S3). As a result, at most $10,000$ data points were kept for each sample; samples from different platforms could then be treated homogeneously by the step-fitting algorithm in downstream analysis.

The 4 features, i.e. S_{peak}, l, v and σ, as defined in the Reliability assessment metrics in the Results and discussion section of the main article, were then extracted for the preprocessed samples. The first 3 features were log-transformed to correct for skewness and then all 4 features were standardized (which means the resulted features have mean 0, standard deviation 1). The pairwise scatter plots of the 4 features are shown in Figure S4. A 3D visualization of the 4D features is shown in Figure S5, which demonstrates that the two classes (reliable/unreliable) represented by the 4 features are highly separable. Note that the principal component analysis (PCA) was adopted for the purpose of visualization only (it is a common way to visualize 4D in a 3D world where we are living); The input of the SVM is still 4 dimensional.

Building custom training set

We recommend that when building custom training set one should first consider adding their own training data into the dataset provided and retrain the model with all the data.

In the cases where this would not be suitable, given that our SVM classifier only uses 4 features we believe 100 samples should usually be enough to obtain a useful classifier. In practice, users should monitor this by splitting their data into a training set and a validation set, and train the model with subsets of different sizes of the training set, and track the prediction performance on the validation set. If for example, the performance on the validation set stops improving as soon as we are training with more than 50% of the data available, we would know that we have more than enough data to train the model. However, if the performance with 95% of the training data is a lot better than the performance with 85% then the model would almost certainly benefit from more training data.

References

1. Zhang, L., Zhang, L.: Use of autocorrelation scanning in DNA copy number analysis. Bioinformatics **29**(21), 2678–2682 (2013)

2. Cai, H., Kumar, N., Baudis, M.: arrayMap: a reference resource for genomic copy number imbalances in human malignancies. PLoS One **7**(5), 36944 (2012)

3. Cai, H., Gupta, S., Rath, P., Ai, N., Baudis, M.: arrayMap 2014: an updated cancer genome resource. Nucleic acids research, 1123 (2014)

4. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: NCBI GEO: archive for functional genomics data sets–update. Nucleic acids research **41**(D1), 991–995 (2013)

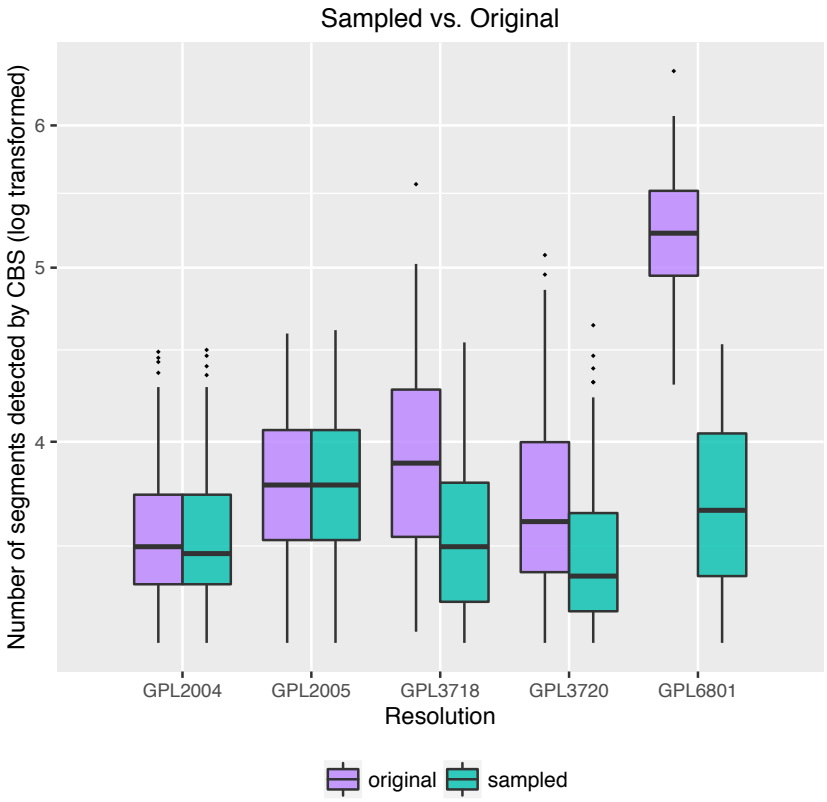


Figure S1: Boxplot of the number of change-points detected by CBS for sampled data (green) versus original data without sampling (purple) for platforms of different resolution. GPL2004 and GPL2005 have resolution of the order of 50,000; GPL3718 and GPL3720 have resolution of the order of 200,000; GPL6801 has resolution of the order of 1,800,000. The range of the number of change-points detected by CBS for sampled data are more consistent among platforms of different resolution than original data without sampling.

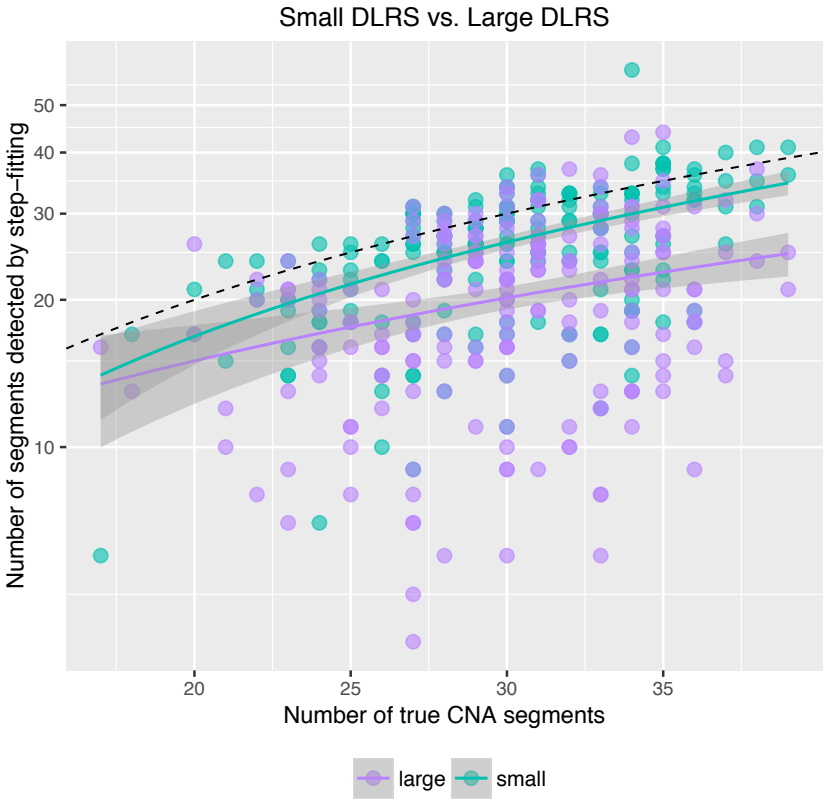


Figure S2: Simulation of 200 set of copy number profiles with small DLRS and large DLRS containing the same true CNA segments of 10,000 dimension. CBS on samples with small DLRS (green, Spearman's $\rho = 0.51$) outperforms that on large DLRS (purple, Spearman's $\rho = 0.29$).

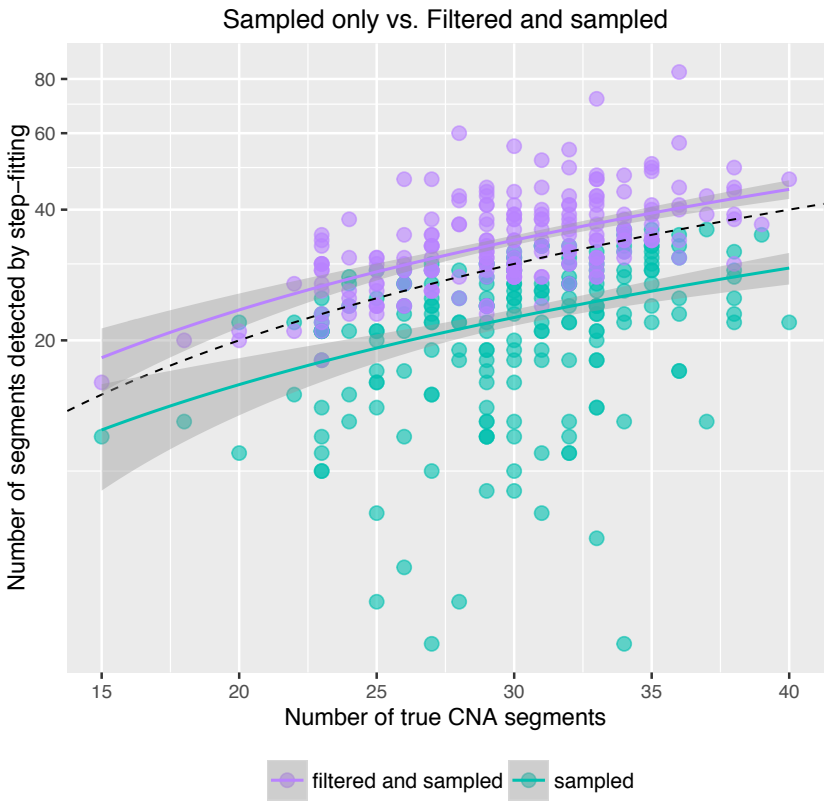


Figure S3: Simulation of 200 set of copy number profiles through different preprocessing procedure. The dimension of the samples is originally 200,000. Green shows the sampled only procedure and purple shows filtered and sampled, where purple (Spearman's $\rho = 0.59$) outperforms green (Spearman's $\rho = 0.41$). The resulted samples are of 10,000 dimension for both procedure.

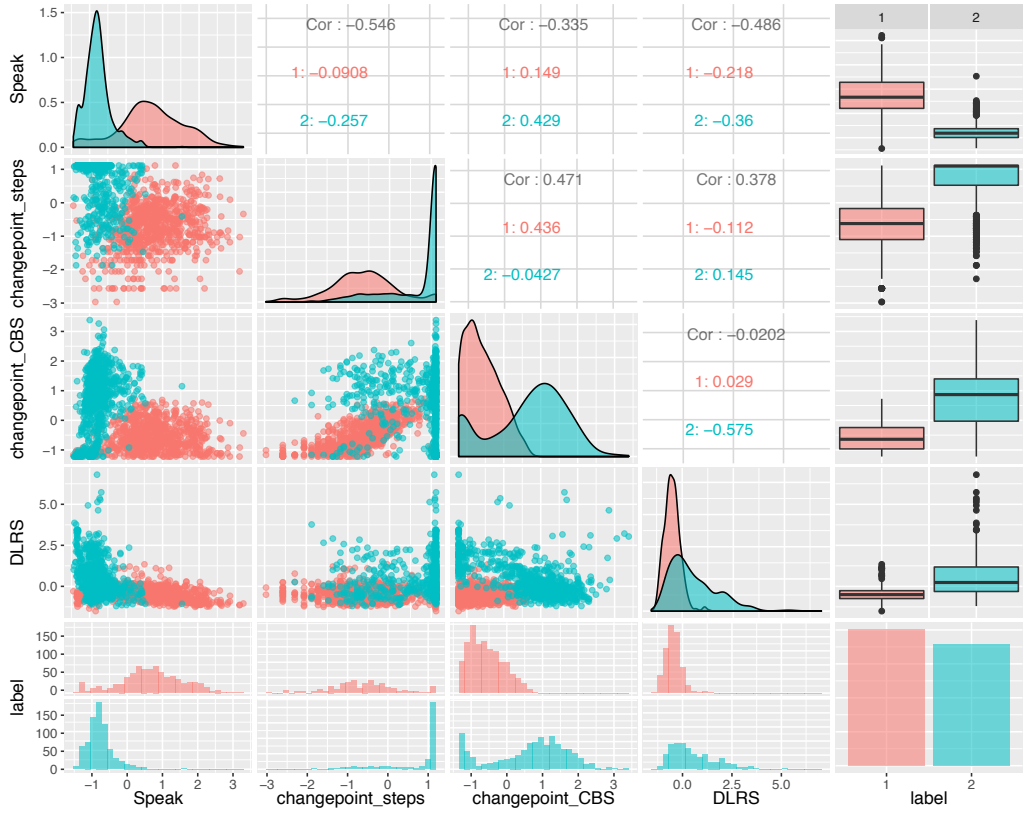


Figure S4: Pairs plot of the 4 features (log transformed and standardized). Red (label 1) denotes reliable; green (label 2) denotes unreliable.

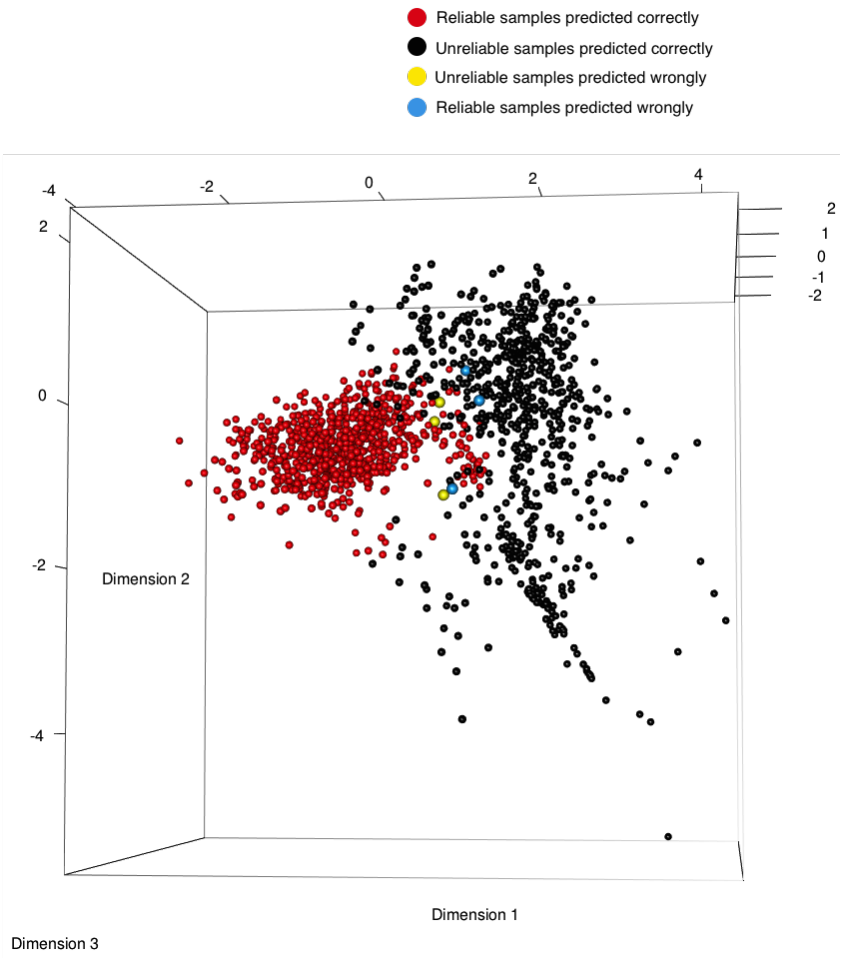


Figure S5: 3D visualization of the 4D dataset (1522 samples, Supplementary Table S1) where dimension reduction was achieved by principal component analysis (PCA) for the purpose of visualization only. Red are correctly predicted reliable samples; Black are correctly predicted unreliable samples; Yellow are unreliable samples predicted as reliable; Blue are reliable samples predicted as unreliable.